中文科技论文图表摘要设计研究*

——以图书情报领域为例

包楚晗[1] 贾丹萍[1] 何 琳[1,2] 马晓雯[1] 艾毓茜[1]

[1](南京农业大学信息科技学院 南京 210095)

[2](南京农业大学领域知识关联研究中心 南京 210095)

摘要:【目的】探究与设计基于图书情报领域、中文科技论文图表摘要构建的结构, 并制定构建规则。【方法】通过调研的方法, 结合人工标注结果及图情领域中文科技论文、图表的特征, 设计摘要框架并规定构建规则, 最终设计评测系统, 基于 SPSS 统计结果分析揭示该摘要系统的表现。【结果】本研究构建的图表摘要在图片信息理解程度、效率、确信度等维度上的表现均优于现有图片–文本组合模式。【局限】图片信息覆盖率有待提高、未考虑清楚图表类型所带来的差异、未完全实施自动化标引。【结论】依据本研究设计的中文科技论文图表摘要构建结构与规则所形成的图表摘要能有效提高用户对文献主要内容的准确理解度。

关键词: 图表标引 中文摘要 李克特量表

分类号: G25

DOI: 10.11925/infotech.2096-3467.2017.0491

1 引 言

在科研过程中, 用户经常需要浏览大量的科技论文以获知领域发展情况或了解相关专业知识, 然而通常情况下, 数据库的检索结果可达数百篇, 对于某些热门且发展较成熟的领域而言, 甚至可达上万篇。在这样的形势下, 用户通过阅读全文获悉文献的主题内容显然是极其耗时耗力的, 所以笔者认为, 用户一般只会筛选其中某些结构内容进行阅读, 如题名、摘要等, 这也与前期调研的结果相似。然而, 文献作者用来解释文献主题的方式呈现多样化[1], 因此题录信息通常不足以涵盖论文主题, 除却文字之外, 图表是最经常被用于佐证研究结果的形式[2], 其中包含大量论文主题的关联信息[3]。因而, 若能够构造精简且包含图表内容及其所揭示内容的图表摘要, 对于科研人员而言, 无疑是帮助其更高效理解文献主题最有力的途径。

国内外已有不少学者对图表标引做了相关研究, 但其处理对象大多为英文文本, 基于汉语与西方语言体系之间的极大差异, 目前的研究结果对于中文文本而言存在一定的不适用性, 且大多是对基于算法模型建立的特征进行自动抽取, 忽略了用户在科研过程中获取知识这一系列行为的特点。

因此, 本研究主要针对以下三点做出改进: 通过实地访谈的方式, 从用户在科研活动中的行为习惯这一角度, 形成特有的图表摘要抽取方式; 基于中文科技论文进行图表摘要研究, 从一定程度上改进了现有研究对象大多为英文这一现象的不足; 结合用户科研习惯, 生成了一套合乎逻辑的摘要组织方式, 更有利于用户获取科技知识。根据本研究构建的摘要组织方式而形成的图表摘要, 能够为用户提供一种解释论文主题的新方式。

2 研究现状

文献中与图表关联的信息极其分散, 图表摘要的

通讯作者: 何琳, ORCID: 0000-0002-4207-3588, E-mail: helin@njau.edu.cn。

*本文系南京农业大学 SRT 计划基金项目“基于自然语言理解的科技论文图表自动标引研究——以生物医学领域疾病研究为例”(项目编号: 201610307061)的研究成果之一。

目的是将这些信息中与文献主题内容关联性较大的部分抽取出来并基于规则合成一段简要且符合逻辑的文摘，以助于用户对图表信息及文献主题信息的攫取。图表自动摘要(Figure Summarization)[4]这一概念自提出以来就受到了极大关注，经笔者梳理发现，目前学者主要针对特征抽取、判定句子权重、建立标引算法、摘要组织、确定评价标准这 5 个关键技术展开大量研究。

2.1 特征抽取

用于抽取图表摘要的特征主要分为以下三类:

(1) 物理特征

物理特征包括句子的位置特征、句子的长度特征、词频特征三种。Luhn[5]提出高频词更有利于揭示论文主题; Nakov 等[6]指出文章中引用句子周围的文本更为重要; 周浪等[7]则提出基于词频分布变化统计的关键术语抽取方法。

(2) 语义特征

语义特征包括与图表、重要段落、题名三种主体有关的语义相似度，可选用 VSM、simHash、LSA (Latent Semantic Analysis)等模型算法对语义相似度进行计算。如 Hirao 等[8]利用句子间相似性的特征判断句子的重要程度; 张帆等[9]利用关键词词表及领域词表对文章题名进行处理，并将文章中与处理后的题名相似度高的语句作为与主题相关度高的对象抽取出来。

(3) 文本特征

文本特征包括相关句的句法特征、相关段落的结构特征、关键词语三方面。如 Brunn 等[10]利用实体词间的联系进行句子抽取; 王芳等[11]研究了 2000 年–2013 年《情报学报》上刊登文献的语法结构等特征，发现句子中心语部分一般由“理论”、“模型”等词语充当，而理论本身通常是句子的定中短语(或称偏正短语); Dahl[12]和 Parkinson[13]在文献创新点抽取过程中，运用语言学特征总结区分了 7 类重要特征的引导词例。

现有抽取特征维度众多，但大多数抽取模型偏重于就其中一种维度进行研究，且大多特征建立在算法统计的基础上，较少注意到用户本身的科研需求。

2.2 句子权重计算

计算句子权重是通过某一种测度方法将句子与图表的关联程度量化，以选择出那些更加重要的句子作为摘要组成。现有权重判定的观点主要分为以下几种:

(1) 基于物理特征

以下两种观点被普遍接受: 段落开头和结尾的句子更能解释论文主题; 可以以具有提示性词语如“如图 X 所示”的句子为中心，以滑动窗口 n 为界限，前后截断(1+2n)个句子作为关键句。

(2) 基于语义特征

利用余弦相似性等算法计算句子与图表标题、重要段落、题名的语义相似度，认为分值越高则关联性越大。如 FigSum+算法[14]即通过计算所抽取句子与图表标题的 TF-IDF 值来筛选得分较高的句子; Ranking SVM[15]函数是一项基于 PairWise 方法的机器学习算法，其实质是利用 SVM 为每一个句子赋予一个分数，以此作为判断其与中心句关联程度的依据，即分数越高，其与中心句的关联越紧密; 潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)模型是基于语义分析的模型，一般会同时利用贝叶斯算法进行关联性高的句子的抽取操作[16]。

(3) 基于文本特征

句子与中心句距离越近，则越重要。如 Radev 等[17]基于质心对句子进行聚类，从而抽取出距离质心最近的句子和它周围的句子。

2.3 标引算法设计

目前，研究人员设计的图表内容自动摘要算法，主要分为两种:

(1) 非监督式标引算法: 即不受文献既定的约束条件，直接依照某种方法进行图表相关文本抽取的方式。例如，从文本中随机抽取 n 个段落，将这些段落的首句经整合直接作为图表的摘要。但非监督式标引算法下分的几种实际操作方法大多具有随机性较大的特点，所以用此算法形成的摘要在完整性和准确性上表现均较差。

(2) 监督式标引算法: 即图表相关文本抽取的过程会受到文献甚至图表内容既定的条件约束，除此之外，抽取的过程通常结合多种抽取特征，并将经权重判断、筛选后的 n 个句子作为图表摘要的预选内容。例如，FigSum+算法[14]综合了: 相似度(Similarity)、TF-IDF 值、表面线索(Surface Cue)、段落特征(Paragraph)、混合特征(Hybrid)共 5 种特征以确定符合抽取要求的句子; FigSum 算法[18]则将全文的句子分为前言(Introduction)、方法(Methods)、结果(Results)和讨

论(Discussion) 4 个部分，而后分别在这 4 个部分中抽取与图表标题语义最为相近的 m 个句子，最后将其组合成结构化的摘要——分别对应于图表的：背景、分析方法、研究结论或成果、揭示的意义。

2.4 摘要组织方式

将筛选出的句子按照一定的方式组成一篇完整的摘要，目前的组织方式可分为两类：抽取式和生成式。

抽取式摘要组织方式仅将已筛选出的句子做简单的连接，而不做其他调整；生成式摘要组织方式则是基于系统已有的专业领域语料库，通过自然语言处理的方法，构建新的语义相似的句子以取代筛选出的内容。如 FigSum[18]系统中按照背景、实验方法、研究结果和研究意义组成图表摘要。

2.5 摘要系统评价

目前对于图表自动摘要系统的评价方法可分为两种：

(1) 直接评价：直接对摘要系统生成的文摘做内容分析，通过与其他模型作比较判断其流畅程度、内容完整度。

(2) 间接评价：通过评价模型对某一任务的完成情况评测该系统，比如根据生成结果的查全率(Recall)、查准率(Precision)和 F 值等。且研究者根据不同任务需求和特点，已生成融合了 P 值、R 值等多种指标在内的综合型评定方法，如关鹏等[16]提出的 LDA 科技文献主题抽取效果评价体系就是融合了包括 P 值、F 值等定量方法及基于主题抽取的广度和主题粒度的定性评定方法的综合型评价体系。

由此，依据现有研究成果的不足之处，本文结合文献及访谈调研的结果，基于用户对文献主题的理解，对中文科技论文图表摘要的设计展开研究，构建一套针对中文科技论文的图表摘要规则，以帮助用户更高效地理解科技论文的主题。

3 论文图表摘要结构设计与构建

为了构建中文科技论文图表摘要，本研究设计了以下研究流程，如图 1 所示。

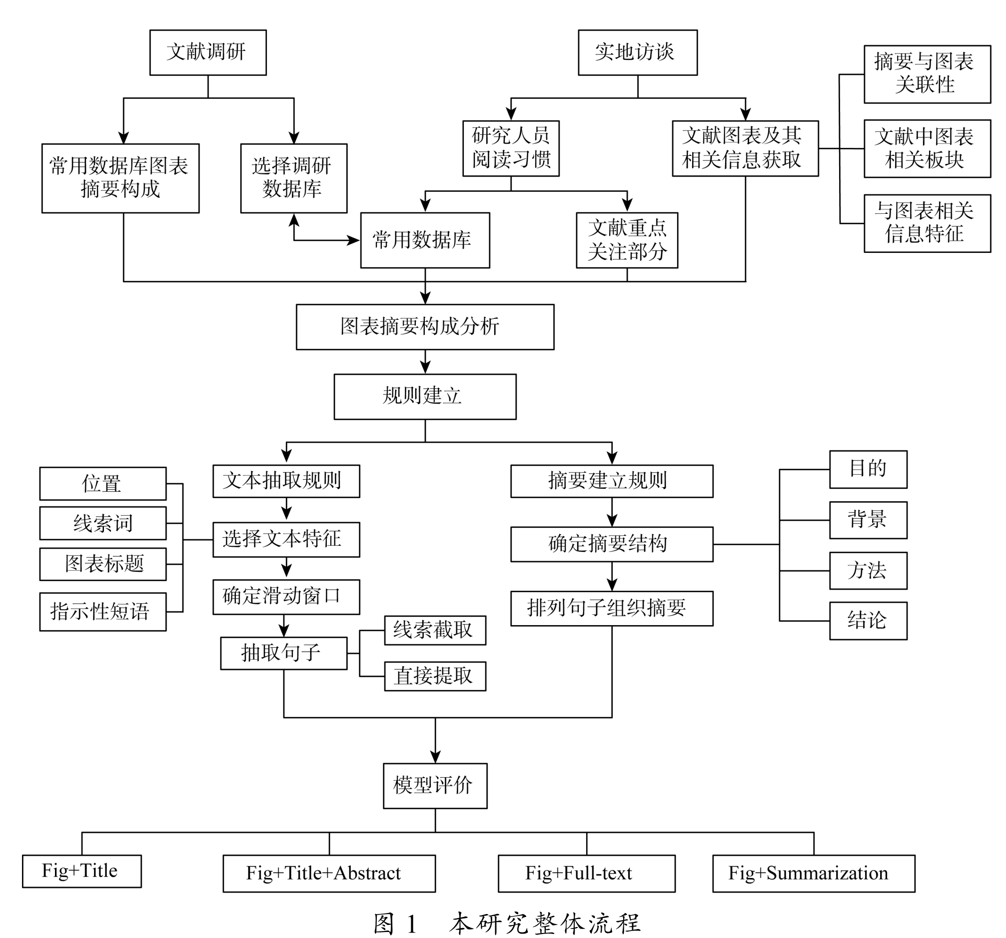


图 1 本研究整体流程

本研究将从文献及访谈调研入手，总结得出基于用户研究需求的图表摘要应当具备的组成部分，依照访谈得出的文本位置特征和结构特征，采用人工标注的方式对一定量的论文进行处理，通过大量的规律总结确定抽取图表摘要的标准和规范，建立用于构建图表摘要的一系列规则，包含图表摘要信息的抽取规则与组织规则。除此之外，还设计了一套评测系统，包含 4 种图表–文本组合模式，通过对评测结果的剖析，对科技论文图表摘要的相关研究起到一定的推动作用。

3.1 摘要结构的设计

对于文献的理解少不了用户的主观判断，这个过程涉及许多难以量化的判断，比如用户的学术思想、个人经验等。因此，本研究对领域共 20 名硕士研究生及博士研究生进行了半结构化面对面深度访谈，通过此次调研，了解了以用户的实际经验来看，一篇完整的图表摘要应具备的基本构件，基于此，探寻各构件与文献文本结构间的关联并分析探讨了每一构件通常存在的位置及其特征等。

本研究未选用通常大样本调查采取的问卷调查方法，而是采用小样本的面对面深度访谈和文献资料查阅，之所以如此，主要是为了避免由于被调查者在填写问卷时为赶时间而敷衍了事所造成的偏差导致无实际意义的情况，并且问卷调查的方式会产生答案的封闭性，若用此方法采集科研人员对于类似本研究中极面向思维的问题的答案就会导致研究人员无法获悉被调查者的真实思维过程，最终得出的结论也只是笔者及团队人员的推断，因此，设计出的问卷的信度与效度都会较低，同时问卷调查的回收率与质量都难以保证，面对面深度访谈虽然需要更大工作量的前期准备与后期整理，也在一定程度上加大了调研的实施难度，但基本能够保证信息与资料来源的客观性、开放性并提高了调研的信度与意义。

访谈提纲的设计如图 2 所示，以了解图表摘要的基本结构为基础，访谈的形式包括面对面访谈与线上约谈(均为个人)两种方式。

由于本项调研在方法选择上的特殊性，获得了大量内容丰富却相对繁杂、不易处理的信息，通过梳理，汇总分析后形成以下几项结果，虽不是访谈结果的全部，但笔者认为是重要且对后续研究极具有现实意义的内容。

序号	题项
1	在什么数据库检索、下载文献?
2	如何判断文献内容是否符合自己需求?
3	文献哪些内容能够帮助快速理解全文?
4	阅读图表时哪些内容是重要的、有助于理解的?
5	文献中与图表相关且重要的信息的特征?
6	文献图表与文献摘要所揭示内容的关系?
7	与图表相关的哪些信息有助于快速理解全文?

图 2 半结构化面对面深度访谈提纲

(1) 领域中研究人员常用的数据库有万方、知网、WOS、谷歌学术、百度学术、Springer、影响因子较高的学术期刊、PubMed、NCBI、日本生物信息统计网站等。

(2) 用户在数据库中检索文献时，经常通过标题、摘要、关键词、期刊杂志、图表(横纵坐标)、引用文献与发表年份等文献构成要素判断文献是否为自己所需(要素排列按统计票数由高到低排列)；研究人员在快速阅读一篇文献时为准确理解文献主题内容会着重阅读文献的摘要、结果、讨论等部分，其次是引言、方法及图表图释部分。

本研究据此提出假设，研究人员在依据经验浏览全文文献并判断文献是否可为自己所用时的依据与文献主要内容有密切的联系。

(3) 研究人员认为在阅读文献时，文献中的图表对于快速了解文献主题内容有极大帮助。他们认为文献中的图表对于实际实验研究有重要的辅助参考性，有时图释信息还会写明实验的简要操作过程。有被访者提到，专业且经验充足的学者在阅读英文文献时，通常会在摘要后首先看图表图释，因为外文文献的研究已经初步形成了基于全文文献的图释，所以这些学者可以根据图释及本身的研究经验了解文献研究的主要内容。而对于中文文献而言，因为其图表缺少类似的图释信息，学者在看到图表之后还需要文献中其他部分的信息才能够大体了解文献的主要内容，这说明图表摘要对于快速准确地理解全文内容起到重要的作用，也在一定程度上佐证了本研究的实际意义。

文献及访谈调研的梳理流程如图 3 所示。根据调研的结果，本文提出，中文科技论文图表摘要的结构应包含以下部件：目的、背景、方法及结论。

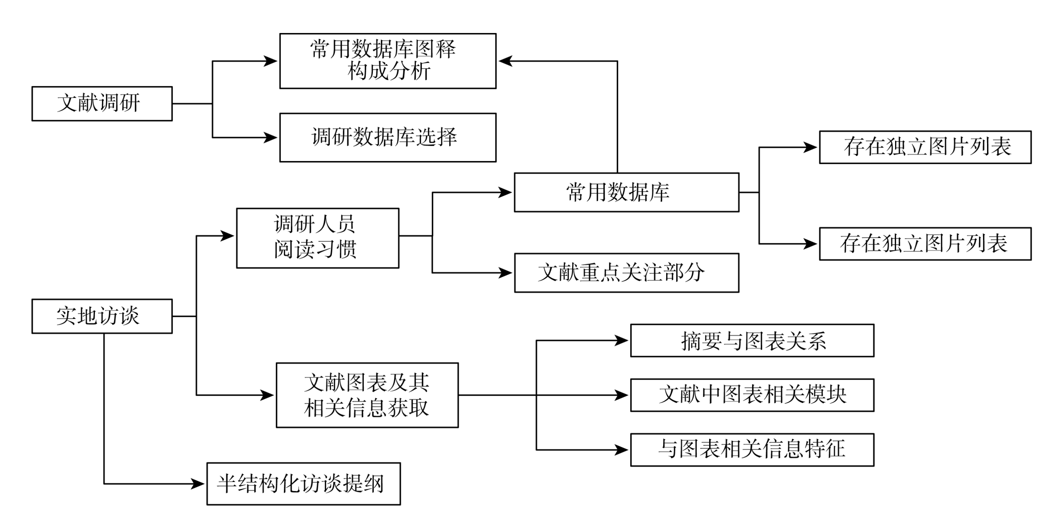


图 3　文献及访谈调研的梳理

3.2　摘要抽取的方法

通过访谈及文献调研发现，论文中与图表相关的信息分散在文本中的各个部分，图表标引就是要找出这些分散的有效信息并进行集中，即将文献文本中与图表内容最相关的若干句子抽取出来并整合成一段简短的结构化摘要，帮助用户理解图表信息，进而提高对文献主题内容的理解。通过分析图表标引方法，本文认为图表标引的关键技术主要包括特征选取、判定句子权重、建立标引算法、组织摘要、确定评价标准 5 个方面。本文所用技术如图 4 所示。

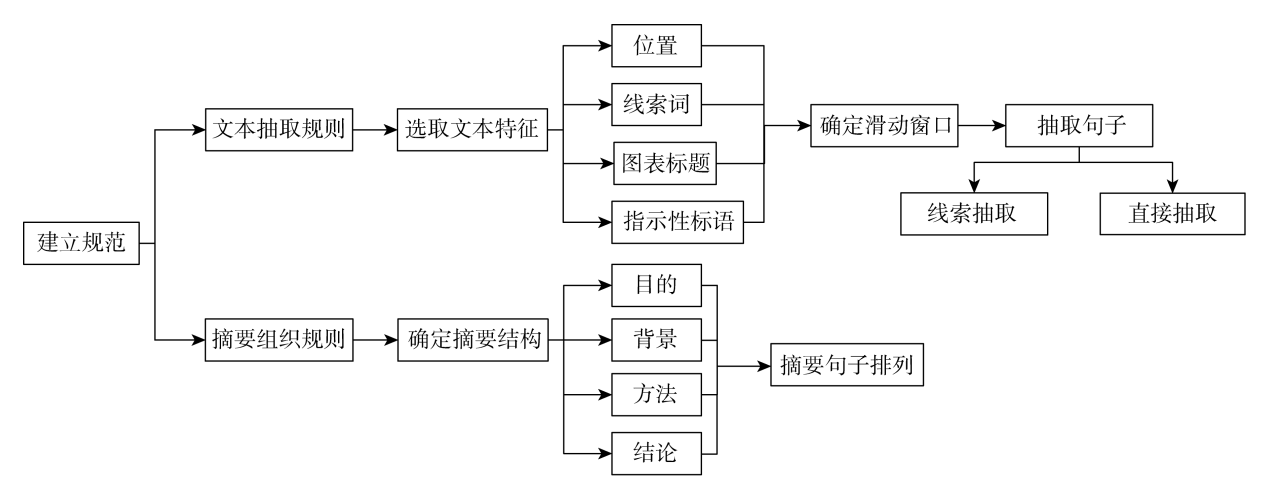


图 4　中文科技论文图表摘要抽取方法

(1) 特征选取

在综合考量已有的理论研究和实践成果后，本文抽取特征有文本物理特征，如句子的位置特征、长度特征及与图表相关的文本特征。

除此之外，本研究认为还可根据文本特征进行句子权重的判定，即根据上述列出的特征可以判定文献中的句子对于揭示图表内容的重要性，并赋予每个句子分值，最后将权值较高的句子抽取出来作为关键句组成摘要。

利用文本物理特征判定权重:

①论文的开头揭示了论文的中心思想和观点，句子与论文开头的距离越近，与论文主题的相关度越高；论文的结尾概括了论文的主要成果和结论，句子与论文结尾的距离越近，与研究结果的相关度越高。通常段首句和段末句也被赋予较高的权值。

②与提及图表的文本的距离，直接影响到该句子与图表内容的相关度；以提及图表的句子为中心距设定一个滑

动窗口的范围，落在滑动窗口之外的句子与图表内容关联度不高。

另外文献中包含一些指示性的词语如：表明、反映、说明等，或者是一些指示性的短语如："如图 X 所示"等，这些词或短语可以作为找到与图表内容相关的文本的线索特征。根据这些线索词和指示性短语可以找到与图表相关的关键句和特征段落。

综合考虑上述文本特征，得分最高的 n 个句子最终被抽取出来作为图表摘要的内容。

(2) 抽取规则的建立

本研究应用抽取式的方法，即从文献文本中抽取出相关度最高、能提供有用信息的句子，并将其联接起来构成一段摘要用以标引图表；而摘要式的方法则需要先理解文献的中心思想，基于这些思想和文献主题，运用自然语言处理的方法构建新的高度概括的句子以取代文献文本中相同内容但较为繁杂的句子，并整合生成一段摘要用以标引图表。本文围绕如下几方面制定了用于处理摘要原始数据的标准。

首先是抽取规则，将其分为两个部分：

①直接提取

1)标题：即图表的名称，如果图表下方无注明，则利用关键词作为线索，提取文献中记录的图表名称。

2)注释：即图表下方的注释，若无，则略过。

②线索截取

1)线索词[19]：根据线索词，如："说明"、"表示"、"可以看出"等，找出文献文本中相应结构的图表信息。

2)指示性短语[19]：根据指示性短语，如："如图 X 所示"、"如下图所示"等，抽取出文献文本中有效的图表信息。

3)滑动窗口[20]：即抽取出包括关键句在内的前后 K 个句子作为最终图表信息的原始材料。经过前期人工标注的表现及前人的经验，本文最终规定：K=5。

其次是组织规则，按照 3.1 节设计与构建的摘要结构，即目的、背景、方法和结论，针对每一结构的特征、结合上述抽取规则规定每一结构中包含的图片信息来源及其特征：

①目的：这一部分的图片信息来源于文献文本中的摘要及前言部分。图表摘要的目的主要是为了揭示这张图表要表达的主要内容，通常可以由 1、2 个句子概括完成，但笔者发现有些图表的目的不会在文献文本中单独说明，却和文献目的有极大相似性，所以在抽取过程中，如果无法找到单一的图表摘要目的，可结合文献目的组织成相应的图表目的。

②背景：这一部分的图片信息来源仍是文献文本中的摘要及前言部分。图表背景通常较为复杂，抽取及组织标引都较有难度，通过前期一定量的人工标注发现，图表背景可结合文献背景及学科背景进行组织。

③方法：这一部分的图片信息来源于文献文本中的方法部分。通过前期工作，本文认为，图表方法通常会出现在文献文本中的方法部分，因此，结合上述抽取规则，抽取方法部分的步骤小标题作为图表标引方法部分的信息来源。除此之外，在调研的过程中，发现在自然科学领域中文献的方法部分通常会多次重复该研究引用的领域内某些经典的方法，比如：杜马斯燃烧法。因此可以通过抽取经统计的方法部分词频最高的词语作为图表内容摘要的方法部分。

④结论：在文献图表周边的文字通常难以攫取该图表的结论，因此，本研究认为图表结论部分的图表信息通常来源于文献文本中的结果部分。对于图表标引的结论，本文通过线索词及指示性短语等特征，结合滑动窗口，从文献结果中抽取相应的图表结论，作为图表标引结论部分的信息来源。

3.3 模型的评价方法

(1) 模型指标规定

从前期的文献调研中，笔者认为用户对于图片的理解是可以量化的，通俗而言，即理解了多少信息、用了多少时间、这些信息对于用户而言是否足够他们理解文献的主题内容等。据此，本研究用"信息理解程度"这一指标指代研究人员理解图片信息的程度；用"图片理解效率"这一指标指代研究人员认为自己正确理解图片所用时长；用"信息覆盖率"这一指标指代研究人员认为本研究设计的图表标引摘要覆盖的信息占全文信息的比率。除此之外，还参考了 Yu 等[21]的研究，增设了"确信度"这一变量，即用户本人对自己前三项指标的打分的确定程度，通过用户对 4 种模式下图片信息相关指标打分的确信度评分情况也可以在一定程度上为本研究提供辅助评判 4 种模式优劣的依据。

(2) 评测打分、流程及分析方法

本文设计了一套评测系统，包含 4 种图表–文本组织模式，结合以上 4 项指标，设计了相应的打分表，如表 1 所示。

表 1　评测打分表

图表–文本组合模式	信息理解程度	图片理解效率	信息覆盖率	确信度
图片+标题				
图片+标题+摘要				
图片+全文				
图片+图表摘要				

利用李克特量表(Likert Scale)[22]量化用户对图片的理解能力: 邀请每位参与评测的用户阅读用于评测的文献并理解其文献主题内容, 对4种模式下基于图片及文本信息的信息理解程度、信息理解效率、信息覆盖率及确信度4项指标从1-10分打分(1分是最低分、10分是最高分), 分值的高低预示用户在4种不同模式下理解图片的能力, 进而反映了用户在这4种图表–文本组合模式下, 对文献主题内容的理解程度。

在取得所有参与评测的用户的打分结果后, 在SPSS应用软件中用单因子方差分析法对评分结果进行统计, 然后结合3.2节制定的图表摘要抽取方法对评分结果进行分析。

4　论文图表摘要测评

4.1　评测系统制定

(1) 评测模式制定

综合文献及访谈调研的结果, 本文设计出了4种用于论文图表摘要测评的图表–文本组合模式: 图片+标题(Figure+Title); 图片+标题+摘要(Figure+Title+Abstract); 图片+全文(Figure+Full-text); 图片+(本文构建的)图表摘要(Title+Summarization)。

(2) 评测对象选取

选取30名图书情报专业方向的研究生, 让他们在没有任何时间限制下完成本研究的测评。

(3) 评测图表选择

在中国知网数据库经检索与筛选, 最终选择10篇图书情报领域的、拥有多种图片类型(如: 凝胶图像、表格、事物的图形、模型以及流程图等)、行文也较为典型的期刊论文, 由一名实验人员严格按照第3节制定的规则分别对文献中36幅图示人工生成对应的图表摘要, 并分别对这36幅图示人工抽取了"图片+标题"、"图片+标题+摘要"、"图片+全文"这三种图表–文本组合模式, 模式记录完成之后, 由另一名实验人员按同一步骤检查前一名实验人员是否有主观意识上的偏颇并给予修正。

4.2　评测结果与分析

(1) 文献主题内容理解能力分析

在SPSS软件中, 将信息理解程度、信息理解效率和信息覆盖率按照40% : 40% : 20%的权重分配转换为一个名为"总评"的虚拟变量, 以此作为用户的"文献主题内容理解能力"指标。

分别以"总评"、"确信度"作为因子进行单因子方差分析, 结果如表2所示。

表2　总评与确信度分别作因子的单因子方差分析结果表

因变量	总评					确信度				
源	III 型平方和	df	均方	F	Sig.	III 型平方和	df	均方	F	Sig.
校正模型	1200.014	3	400.005	318.681	0	163.569	3	54.523	20.365	0
截距	57355.779	1	57355.779	45694.903	0	64489.341	1	64489.341	24087.159	0
类型	1200.014	3	400.005	318.681	0	163.569	3	54.523	20.365	0
误差	1501.207	1196	1.255			3202.090	1196	2.677		
总计	60057.000	1200				67855.000	1200			
校正的总计	2701.221	1199				3365.659	1199			

由表2的检验结果可知, 不同图表–文本组合模式下, 用户的图片信息理解能力里与评分的确信度的确会因图表–文本组合模式的不同而有所差异($F_{总评}(3, 1196)=318.681$, $p<0.01$、$F_{确信度}(3, 1196)=20.365$, $p<0.01$)。总评及确信度的均值及偏差如表3所示。

从表3中可知, 随着4种图表–文本组合模式中所提供的文献文本的范围加大, 用户对文献主题内容的理解能力和对评分的确信度也随之增高, 例如, 以"图片+标题"这种图表–文本组合模式为基线, 当增加了文献摘要时, 用户对文献主题的理解能力提高了24.2%, 对评分的确信度也增高了6.54%; 当增加了本文生成的图表摘要时(此摘要中包括图片标题), 用户对文献主题的理解能力提高了41.6%, 对评分的确信度也增高了4.40%; 当提供文献全文时, 用户对文献主题的理解能力提高了48.0%, 对评分的确信度也增高了7.12%。因此, 增加了摘要和标题对用户理解文献主题都有一定程度上的帮助, 但全文内容能够最大限度地增大用户的文献主题的理解能力, 而本文所构建的图表摘较文献摘要表现更好。

表 3　文献主题内容理解能力(总评)与确信度的均值及偏差

图表–文本组合模式	总评	确信度
图片+标题	5.38±1.25	6.88±1.65
图片+标题+摘要	6.68±1.13	7.33±1.58
图片+图表摘要(本研究构建)	7.62±1.18	7.18±1.79
图片+全文	7.96±0.89	7.37±1.68

对图表–文本组合的 4 种模式两两对比后可得到模式间的差异。将用户对文献主题的理解能力作为因子检验后，得到的结果如表 4 所示。

表 4　文献主题内容理解能力因子 F 检验结果

(I)类型	(J)类型	均值差值(I-J)	标准误差	Sig.
图片+标题	图片+标题+摘要	–1.297	0.091	0.001
	图片+全文	–2.239	0.091	0
	图片+图表摘要	–2.580	0.091	0
图片+标题+摘要	图片+标题	1.297	0.091	0.001
	图片+全文	–0.942	0.091	0.006
	图片+图表摘要	–1.283	0.091	0
图片+全文	图片+标题	2.239	0.091	0
	图片+标题+摘要	0.942	0.091	0.006
	图片+全文	–0.341	0.091	1.000
图片+图表摘要	图片+标题	2.580	0.091	0
	图片+标题+摘要	1.283	0.091	0
	图片+全文	0.341	0.091	1.000

由表 4 可知，将用户对文献信息理解的三个维度转换为一个维度时，只有“图片+图表摘要(本研究构建)”与“图片+全文”这两种模式下，用户对文献主题的理解能力无显著差异(显著性 $p=1.000>0.05$)，其余模式之间的比较都呈显著差异(显著性 $p\leqslant0.05$)，由此可知，严格按照本文制定的规则所生成的图表摘要对用户理解文献主题的作用与用户通读全文对理解文献主题的作用呈极大相似性，且其影响程度远大于“图片+标题”模式及“图片+标题+文献摘要”模式。

在得到这些结果的基础上，进一步将用户对文献主题理解的三项指标(即信息理解程度、图片信息理解效率、信息覆盖率)分别作为因子进行分析。

(2) 文献主题内容理解的三项指标分析

①以信息理解程度为因子

利用 SPSS 软件，以信息准确度为单因子进行单因子方差分析，结果如表 5 所示。可知，不同于其余模式成对比较，在“图片+图表摘要”与“图片+全文”这两种模式下，用户对图片信息的理解程度无显著性差异(显著性 $p=0.139>0.05$)，例如在“图片+图表摘要”及“图片+标题+摘要”两种模式下，用户对图片信息的理解程度有显著性差异(显著性 $p=0.061>0.05$)。因此，可以推断本研究所构建的图表摘要对用户而言，对图片信息的理解程度与其通过直接阅读全文文献来理解图片信息的程度有极大相似性，且本研究所构建的图表摘要在这一维度上的表现远好于文献摘要。

结合 3.2 节制定的摘要抽取方法，本研究认为图片信息理解程度的高低与线索词及指示性短语的选取有密切关系，即前期工作中归纳的线索词词表及指示性短语列表是否完整直接影响图片信息的抽取，进而影响用户对图片信息的理解。因此，在今后的研究中，仍需要针对特定领域制定相应完整精确的、有该领域科技论文用词特点的线索词词表及指示性短语列表，以获得更高的信息理解程度。

②以图片信息理解效率为因子

利用 SPSS 软件，以信息理解效率为单因子进行分析，结果如表 5 所示。可知，在“图片+标题”与“图片+全文”这两种模式下，用户对图片信息的理解效率无显著性差异($p=1.000>0.05$)，在“图片+标题”这一模式下，虽然用户用于理解图片的速率最高但其理解图片信息的效果最差，而在“图片+全文”这一模式下，虽然用户理解图片信息的效果最好但其用于理解图片的速率最低，所以两项指标结合使得两种模式下，用户对于图片信息的理解效率都较低；在“图片+标题+摘要”与“图片+全文”两种模式下，用户对图片信息的理解效率无显著性差异($p=0.639>0.05$)，相较于“图片+全文”，在“图片+标题+摘要”这种模式下，用户用于理解图片的速率有一定程度下降但其理解图片信息的效果有一定程度的提高；在“图片+图表摘要”与“图片+全文”两种模式下，用户对图片信息的理解效率有显著性差异($p<0.05$)，结合上文分析以及在“图片+图表摘要”与“图片+标题+摘要”两种模式下，科研人员对图片信息的理解效率的无显著性差异($p=0.029<0.05$)，笔者认为，本研究构建的图表摘要对用户理解图片信息而言，效果虽不如全文文献模式但较文献摘要模式而言更好，速率虽不敌文献摘要模式但较全文文献模式好。总之，在这一维度上，本研究构建的图表摘要的表现较其余三种模式更优。

结合 3.2 节制定的摘要抽取方法，本文认为图片信息理解效率与滑动窗口 K 的选取有密切关系。图片信息理解效率较高，这与图表摘要的篇幅及信息理解程度指标的表现均有关联。对于篇幅而言，滑动窗口 K 的选取也并不是越小越好，而是要结合较高程度的图片信息理解能力综合评判，因此，在今后的研究中，需要选取不同的 K 值进行试验，通过比较试验结果中图片信息理解程度的高低来确定较为合适的 K 值。

③以图片信息覆盖率为因子

利用 SPSS 软件，以信息覆盖率为单因子进行分析，结果如表 5 所示。可知，在这一维度下，4 种模式成对比较之后，

4 种模式下的图片信息覆盖率均有显著性差异(p≤0.05)，笔者对此有两种推断：本研究构建的图表摘要中所含有的图片信息比文献摘要或全文文献中的图片信息更专指、对用户理解图片更能起作用；本研究所构建的图表摘要包含较多对用户理解图片起作用的信息，但与图片关联却对用户理解图片作用不大的信息的覆盖率较低。

结合 3.2 节制定的摘要抽取方法，本研究认为信息覆盖率的高低与滑动窗口 K 的选取有密切关系，K 值的选取偏大或偏小都会导致信息覆盖率偏低。因此，在今后的研究中，需要根据学科领域科技论文的造句特点选取不同的 K 值进行试验，通过比较试验结果中信息覆盖率的高低确定较为合适的 K 值。

表 5　基于三项指标的模式成对对比

模式		信息理解程度			信息理解效率			信息覆盖率		
I 类型	J 类型	均值差值(I-J)	标准无误	Sig.	均值差值(I-J)	标准无误	Sig.	均值差值(I-J)	标准无误	Sig.
图片+标题	图片+标题+摘要	–1.597	0.124	0	–0.937	0.130	0.639	–1.417	0.112	0
	图片+全文	–3.127	0.124	0	–0.323	0.130	1.000	–4.293	0.112	0
	图片+图表摘要	–2.697	0.124	0	–2.260	0.130	0	–2.987	0.112	0
图片+标题+摘要	图片+标题	1.597	0.124	0	0.937	0.130	0.639	1.417	0.112	0
	图片+全文	–1.530	0.124	0	0.613	0.130	0.639	–2.877	0.112	0
	图片+图表摘要	–1.100	0.124	0.061	–1.323	0.130	0.029	–1.570	0.112	0.024
图片+全文	图片+标题	3.127	0.124	0	0.323	0.130	1.000	4.293	0.112	0
	图片+标题+摘要	1.530	0.124	0	–0.613	0.130	0.639	2.877	0.112	0
	图片+图表摘要	0.430	0.124	0.139	–1.937	0.130	0	1.307	0.112	0
图片+图表摘要	图片+标题	2.697	0.124	0	2.260	0.130	0	2.987	0.112	0
	图片+标题+摘要	1.100	0.124	0.061	1.323	0.130	0.029	1.570	0.112	0.024
	图片+全文	–0.430	0.124	0.139	1.937	0.130	0	–1.307	0.112	0

5 结　语

对于科研人员来说，文献中图片下方图释以外的、存在于文献文本中的内容对于图片的正确理解是至关重要的。因此，基于用户的研究需求，通过文献及访谈调研的方法，结合人工标注的结果及图情领域科技论文摘要、图表的特征，本研究以图书情报领域为实证对象，制定了用于构建中文科技论文图表摘要的规则，包括图表摘要的抽取规则及组织规则，并设计了 4 种图表–文本组织模式进行评测。

根据评测结果可知：通过设置合适的 K 值可以提高信息覆盖率；通过建立相应完整精确的、有该领域科技论文用词特点的线索词词表及指示性短语列表，可以提高图片信息的理解程度，进而提高文献主题内容的理解程度；对于提高图片信息理解效率而言，需要在获得较高图片信息理解程度的基础上综合考量滑动窗口 K 的取值。本研究以图书情报为例研究了中文科技论文图表摘要构建的方法流程和实现路径，发现根据本研究设计的结构与规则而形成的图表摘要在图片信息理解程度、信息理解效率及信息覆盖率三大维度上的表现均较优。但也存在不足，在今后的工作中，将就如何确定 K 的取值及如何实现高效地自动化抽取图表摘要等方向进一步研究，并将本文的研究结果在其他学科领域进行实验。

参考文献：

[1] Kim D, Yu H. Figure Text Extraction in Biomedical Literature[J]. PLoS One, 2011, 6(1): e15338.

[2] Yu H, Lee M. Accessing Bioscience Images from Abstract Sentences[J]. Bioinformatics, 2006, 22(14): 547-556.

[3] Agarwal S, Yu H. Figure Summarizer Browser Extensions for PubMed Central[J]. Bioinformatics, 2011, 27(12): 1723-1724.

[4] Futrelle R P. Handling Figures in Document Summarization Abstract [C]//Proceedings of Meeting of the Association for Computational Linguistics. 2004.

[5] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development, 1958, 2(2):

159-165.

[6] Nakov P I, Schwartz A S, Hearst M A. Citances: Citation Sentences for Semantic Analysis of Bioscience Text[C]//Proceedings of the SIGIR'04 Workshop on Search and Discovery in Bioinformatics. 2004.

[7] 周浪，张亮，冯冲，等. 基于词频分布变化统计的术语抽取方法[J].计算机科学，2009，36(5): 177-180. (Zhou Lang, Zhang Liang, Feng Chong, et al. Terminology Extraction Based on Statistical Word Frequency Distribution Variety[J]. Computer Science, 2009, 36(5): 177-180.)

[8] Hirao T, Isozaki H, Maeda E, et al. Extracting Important Sentences with Support Vector Machines[C]//Proceedings of the 19th International Conference on Computational Linguistics. 2002: 1-7.

[9] 张帆，乐小虬. 面向领域科技文献的句子级创新点抽取研究[J]. 现代图书情报技术，2014(9): 15-21. (Zhang Fan, Le Xiaoqiu. Research on Innovation Points Extraction from Scientific Research Paper Based on Field Thesaurus[J].New Technology of Library and Information Service, 2014(9): 15-21.)

[10] Brunn M, Chali Y, Pinchak C. Text Summarization Using Lexical Chains[C]//Proceedings of the Document Understanding Conference, 2001: 135-140.

[11] 王芳，史海燕，纪雪梅. 我国情报学研究中理论的应用：基于《情报学报》的内容分析[J]. 情报学报，2015，34(6): 581-591. (Wang Fang, Shi Haiyan, Ji Xuemei. The Use of Theory in Chinese Information Science Research Based on the Content Analysis of the Journal of the China Society for Scientific and Technical Information[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 581-591.)

[12] Dahl T. Contributing to the Academic Conversation: A Study of New Knowledge Claims in Economics and Linguistics [J]. Journal of Pragmatics, 2008, 40(7): 1184-1201.

[13] Parkinson J. The Discussion Section as Argument: The Language Used to Prove Knowledge Claims [J]. English for Specific Purposes, 2011, 30(3): 164-175.

[14] Ramesh B P, Sethi R J, Yu H. Figure-Associated Text Summarization and Evaluation [J]. PLoS One, 2015, 10(2): e0115671.

[15] Herbrich R, Graepel T, Obermayer K. Support Vector Learning for Ordinal Regression[C]//Proceedings of the 9th International Conference on Artificial Neural Networks. IET, DOI: 10.1049/cp: 19991091.

[16] 关鹏，王曰芬，傅柱. 不同语料下基于 LDA 主题模型的科学文献主题抽取效果分析[J]. 图书情报工作，2016，60(2): 112-121. (Guan Peng, Wang Yuefen, Fu Zhu. Effect Analysis of Scientific Literature Topic Extraction Based on LDA Topic Model with Different Corpus [J]. Library and Information Service, 2016, 60(2): 112-121.)

[17] Radev D R, Jing H, Styś M, et al. Centroid-based Summarization of Multiple Documents [J]. Information Processing & Management, 2004, 40(6): 919-938.

[18] Agarwal S, Yu H. FigSum: Automatically Generating Structured Text Summaries for Figures in Biomedical Literature [C]//Proceedings of AMIA Annual Symposium. 2009.

[19] 朱丽萍，李洪奇，杨中国，等. 一种面向科技文献引言的信息抽取方法[J]. 山东大学学报：理学版，2015，50(7): 23-30, 37. (Zhu Liping, Li Hongqi, Yang Zhongguo, et al. An Information Extraction Method for Scientific Literature Introduction[J]. Journal of Shandong University: Natural Science, 2015, 50(7): 23-30, 37.)

[20] 杜威，邹先霞. 基于数据流的滑动窗口机制的研究[J]. 计算机工程与设计，2005，26(11): 2922-2944. (Du Wei, Zou Xianxia. Research of Sliding Windows Scheme Based on Data Stream[J]. Computer Engineering and Design, 2005, 26(11): 2922-2944.)

[21] Yu H, Agarwal S, Johnston M, et al. Are Figure Legends Sufficient? Evaluating the Contribution of Associated Text to Biomedical Figure Comprehension [J]. Journal of Biomedical Discovery and Collaboration, 2009, 4(1). DOI: 10.1186/1747-5333-4-1.

[22] 方宝. Likert 等级量表调查结果有效性的影响因素探析[J]. 十堰职业技术学院学报，2009，22(2): 25-28. (Fang Bao. An Analysis of the Factors Influencing the Effectiveness of Likert Rating Scale's Investigation Result [J]. Journal of Shiyan Technical Institute, 2009, 22(2): 25-28.)

[23] Lin C Y, Hovy E. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics[C]//Proceedings of the 2003 Conference of North American Chapter of the Association for Computational Linguistics on Human Language. 2003: 71-78.

[24] 傅间莲，陈群秀. 一种新的自动文摘系统评价方法[J]. 计算机工程与应用，2006(18): 176-177. (Fu Jianlian, Chen Qunxiu. A New Evaluation Method for Automatic Text Summarization[J]. Computer Engineering and Applications, 2006(18): 176-177.)

[25] Lin C Y. ROUGE: A Package for Automatic Evaluation of Summaries [C]//Proceedings of the Workshop on Text Summarization Branches out.2004: 74-81.

作者贡献声明:

何琳：提出研究思路，设计研究方案；

包楚晗, 贾丹萍, 马晓雯, 艾毓茜: 进行实验, 采集、清洗和分析数据;
包楚晗, 贾丹萍: 论文起草;
包楚晗, 贾丹萍, 何琳: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: helin@njau.edu.cn。

[1] 包楚晗. diaoyanshenming.docx. 调研前与用户的申明.
[2] 包楚晗. diaoyanhuizong.docx. 调研音频整理分析.
[3] 马晓雯. pingjiabiao.docx. 评测时所用评价表.
[4] 包楚晗, 马晓雯, 贾丹萍. rengongzhaiyao.docx. 基于本研究规则人工生成摘要.
[5] 艾毓茜. pingceyuanshuju.xlsx. 评测后汇总的原数据.
[6] 包楚晗. fenxi.doc. 评测后 SPSS 导出的分析结果.

收稿日期: 2017-05-31
收修改稿日期: 2017-07-11

Summarizing Figures of Chinese Scholarly Articles of Library and Information Science

Bao Chuhan[1] Jia Danping[1] He Lin[1,2] Ma Xiaowen[1] Ai Yuxi[1]
[1](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)
[2](Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: [**Objective**] This paper studies the figures of Chinese articles in the field of library and information science (LIS), aiming to establish new principles to summarize them. [**Methods**] We proposed the framework and rules for figure summarization based on manual indexing and features of LIS papers. Then, we evaluated the performance of the new system with the help of SPSS. [**Results**] Compared with the existing figure-text model, our method could more effectively process information from the figures. [**Limitations**] We need to extract more information from the figures, analyze the influences of different charts, and add automatic indexing functions to the new system. [**Conclusions**] The proposed method could effectively summarize figures from the scholarly articles.

Keywords: Figure Indexing Abstract in Chinese Likert Scale